NOvA Data Acquisition Architecture

M. Bowden, G. Guglielmo, B. Haynes, R. Kwarciany, V. Pavlicek, M.Votava

Introduction

The Data Acquisition (DAQ) system for NOVA is based on a standard Gigabit Ethernet network and commercial processors.

There are 23,808 Front-end Boards (FEBs), each generating approximately 1 Mbps of data. To allow for adjustments in thresholds, inefficiencies in the readout electronics, and some expansion capability, a minimum combined DAQ bandwidth requirement of ~100 Gbps is assumed. The output data rate of a Front-end Board should therefore be at least 4 Mbps.

The detector is assembled as 62 blocks, each containing 32 planes (16 vertical and 16 horizontal) for a total of 1984 planes. A block holds 12,288 detector channels and connects to 384 FEBs.

Data is acquired continuously by the FEBs. Data from the spill window (30 usec per 2 second cycle) is always recorded. Both spill and non-spill data are analyzed by a Processor farm for evidence of a supernova signal. If detected, a 20 second window of data is recorded.

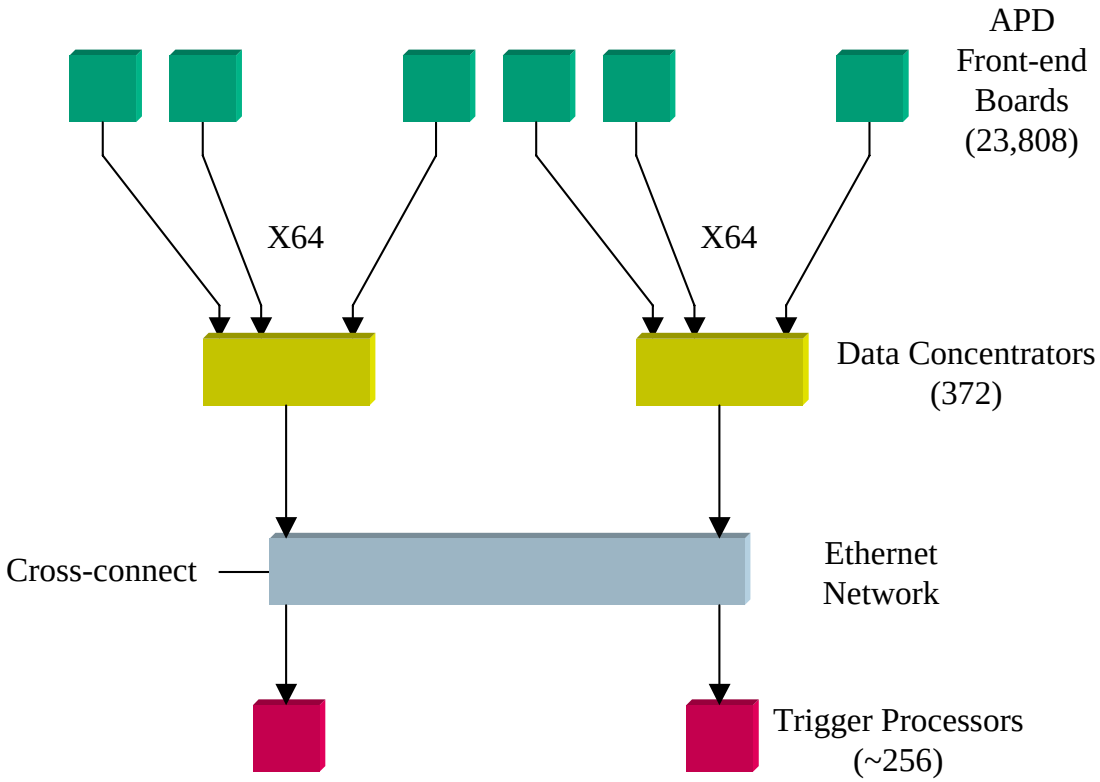

Figure 1. DAQ Architecture

A single Processor receives data from FEBs for a fixed timeslice. The length of a timeslice is somewhat arbitrary, limited mainly by buffer space. A Data Concentrator connects to 64 FEBs (figure 2).

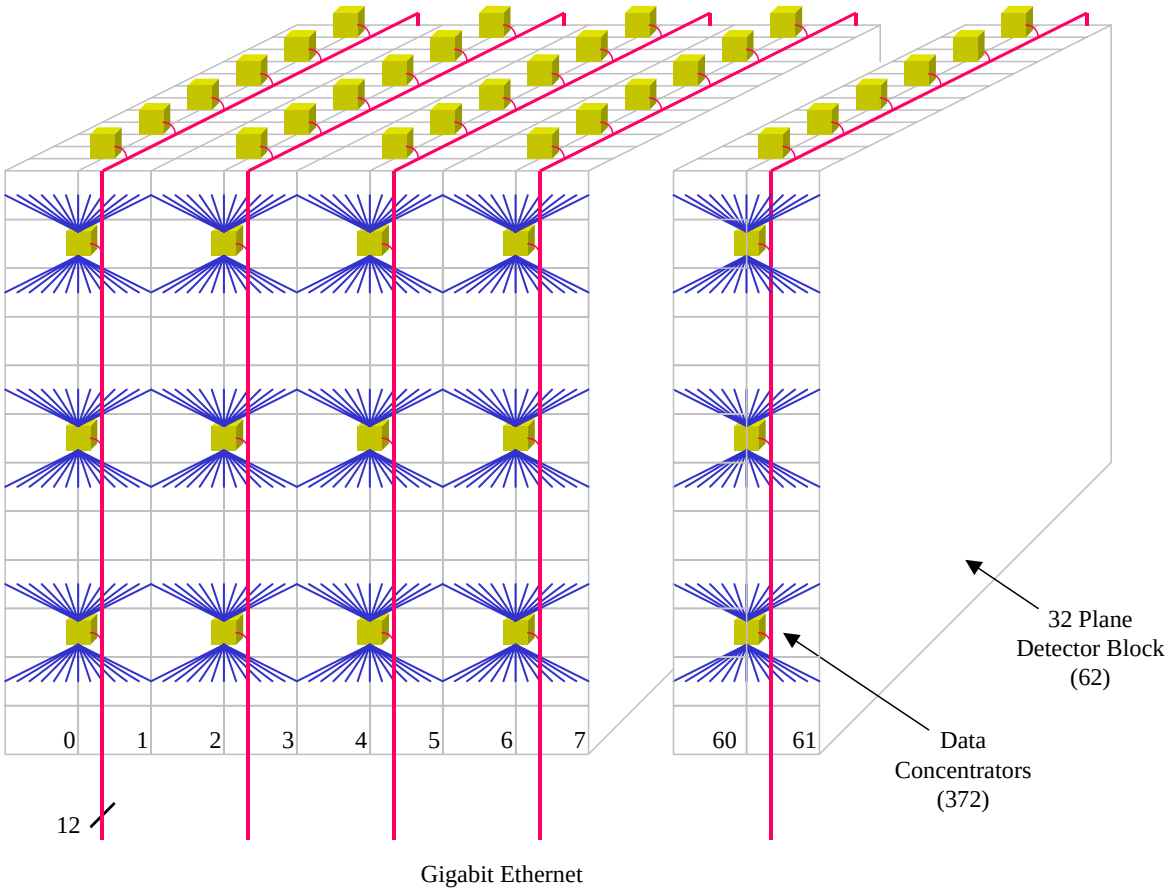

Figure 2. Geographical Arrangement of Data Concentrators

Front-end Interface

Two types of front-end interconnect are under consideration; a "star" configuration similar to most Ethernet installations (Figure 3), and a "loop" configuration similar to a token ring network (Figure 4). The loop configuration reduces cable cost but implies higher rate data transfers.

In both illustrations the Data Concentrator is assumed to be mounted on or near the detector and covers 4 meters in z. Horizontal planes are read out from both sides of the detector, while vertical planes are read out along the top edge only.

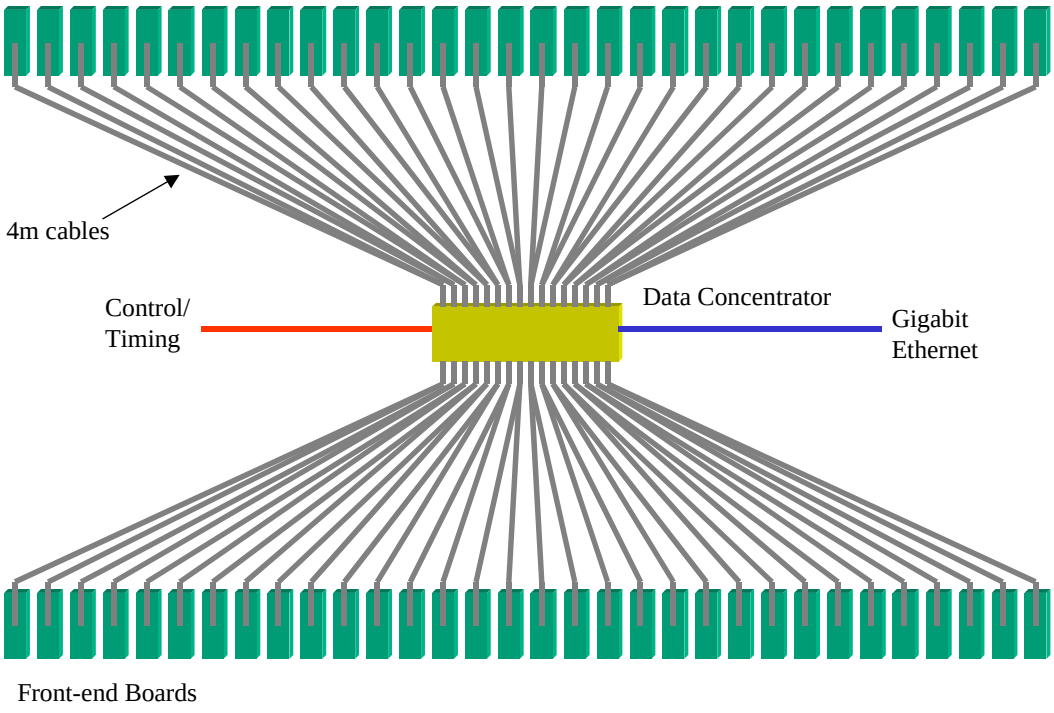

Figure 3. Front-end to Data Concentrator Interface – Star Configuration

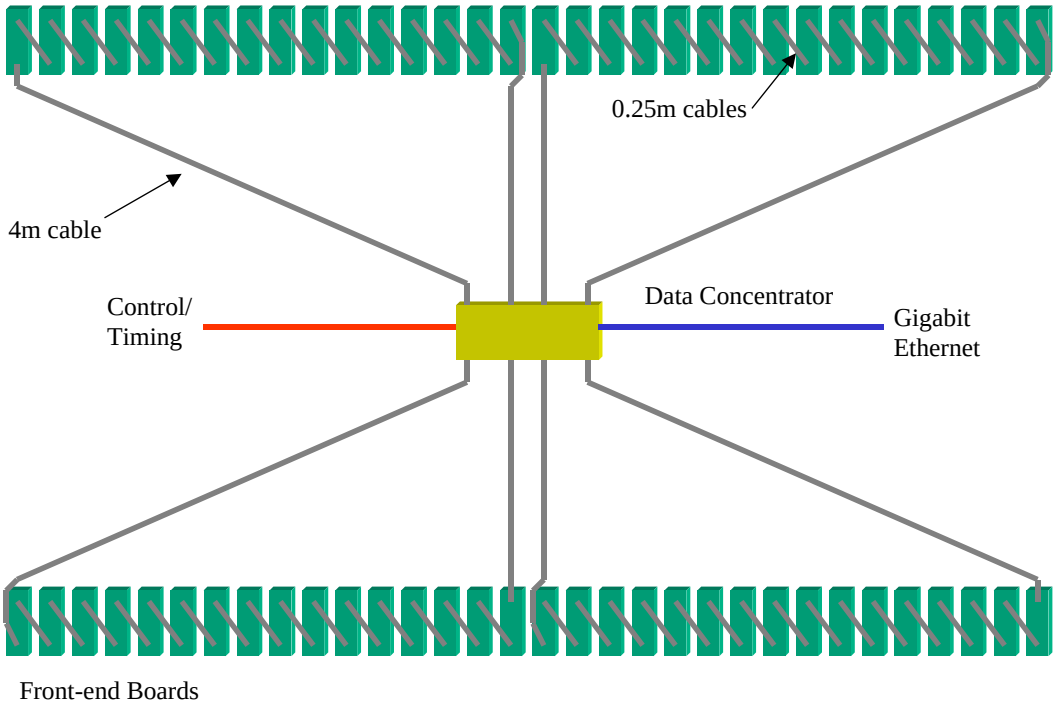

Figure 4. Front-end to Data Concentrator Interface – Loop Configuration

For either configuration, the cable is standard CAT5e with RJ-45 connectors (figure 5). There is a differential clock, and a clock-synchronous "sync" signal used to align all front-end modules to a specific clock. A serial Command channel is used to download and control the front-end modules, and a serial Data channel is used to transmit front-end data back to the Data Concentrator.

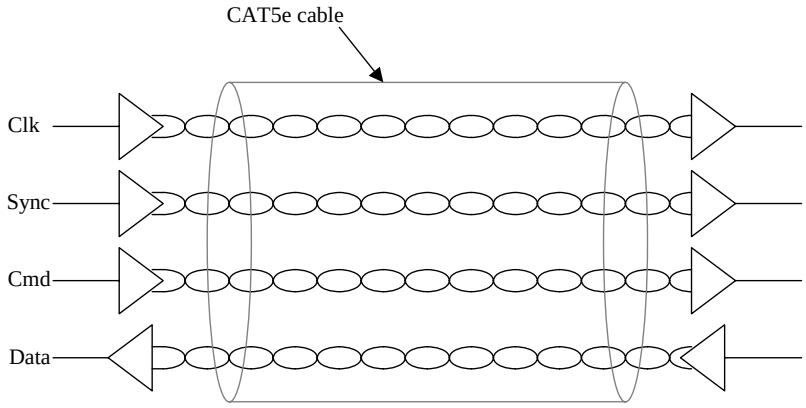

Figure 5. Serial Interface Signals

A simple 4 Mbps serial link using the 16 MHz clock and 4X oversampling is illustrated in figure 6. This is the suggested protocol for the star configuration. The clock is the same at both ends of the link, so there is no restriction on frame size. For the loop configuration, the Data channel must operate at a 16X higher rate, so a faster clock and synchronous protocol is required. The Command channel can be either synchronous or asynchronous.

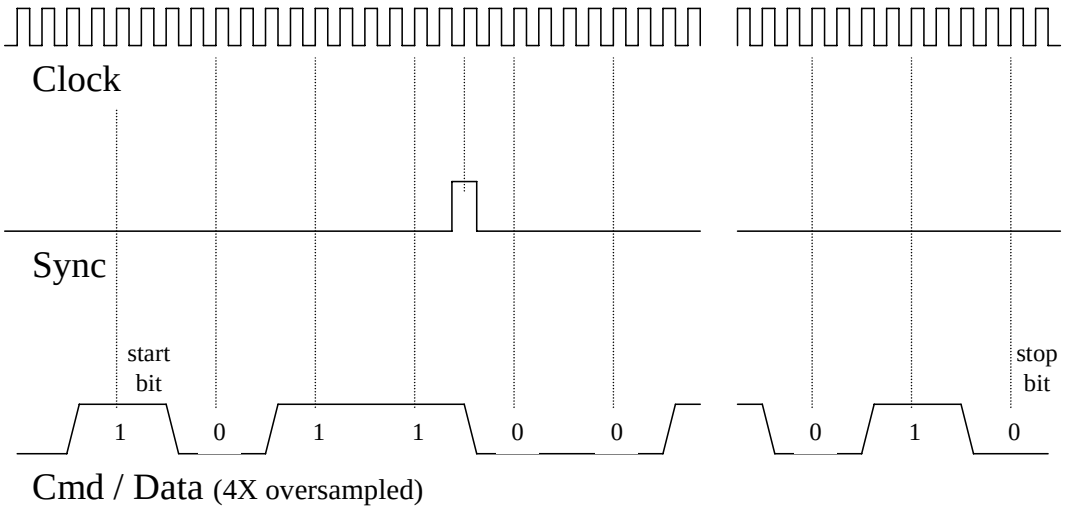

Figure 6. Oversampled Serial Protocol

Issues to be considered in the selection of a front-end interface are;

1) clock synchronization – clock skew (assuming active repeaters) between the first and last FEB in the loop configuration is approximately 50 nanoseconds. If closer synchronization is necessary, a delay element must be included in the FEB. Clock synchronization for the star configuration is provided by equal length cables.
2) data rate – the data transfer rate for the loop configuration is 8-16 times higher, but it allows bandwidth averaging over all FEBs in the loop.
3) reliability – failure of a single FEB in the loop configuration will likely have greater impact.
4) protocol - for both options it is assumed that the signaling is LVDS and the protocol can be implemented directly in an FPGA. The loop configuration requires a synchronous protocol, the star configuration can use either synchronous or asynchronous protocol.
5) cable length – the star configuration requires that the Data Concentrator be placed on or near the detector to minimize cable volume. The loop configuration allows greater separation between Data Concentrator and FEBs, but cable length may still be limited by the higher speed synchronous protocol.
6) additional application – the star version of the Data Concentrator can also serve as a timing system fanout module

Data Concentrator

The Data Concentrator (figure 7) is a small card with 8 (loop) or 64 (star) input ports and a Gigabit Ethernet output port. It is housed in a standalone shielded enclosure resembling a desktop network switch. Depending on pricing, it may contain an FPGA with built-in processor and Ethernet MAC (e.g., Xilinx Virtex-4 FX12), or a lower cost FPGA with external processor/Ethernet. It also contains up to 512 MByte of buffer memory for use in traffic shaping. A block diagram is shown in figure 8.

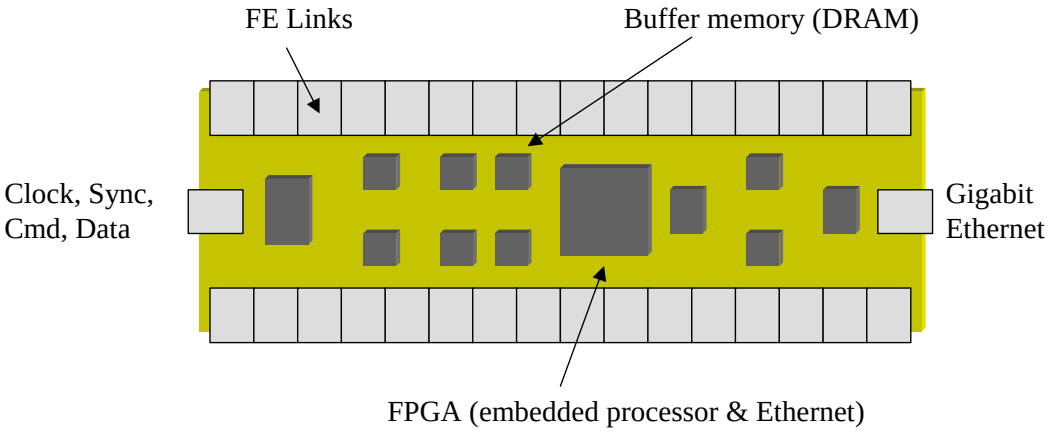

Figure 7. Data Concentrator Board

The expected Data Concentrator output rate is <100 Mbps. A Gigabit Ethernet link can typically support >500 Mbps (up to 900 Mbps if jumbo data frames are enabled). Data Concentrator cost is estimated at $12 per FEB. Development is expected to take 1.5 FTE.

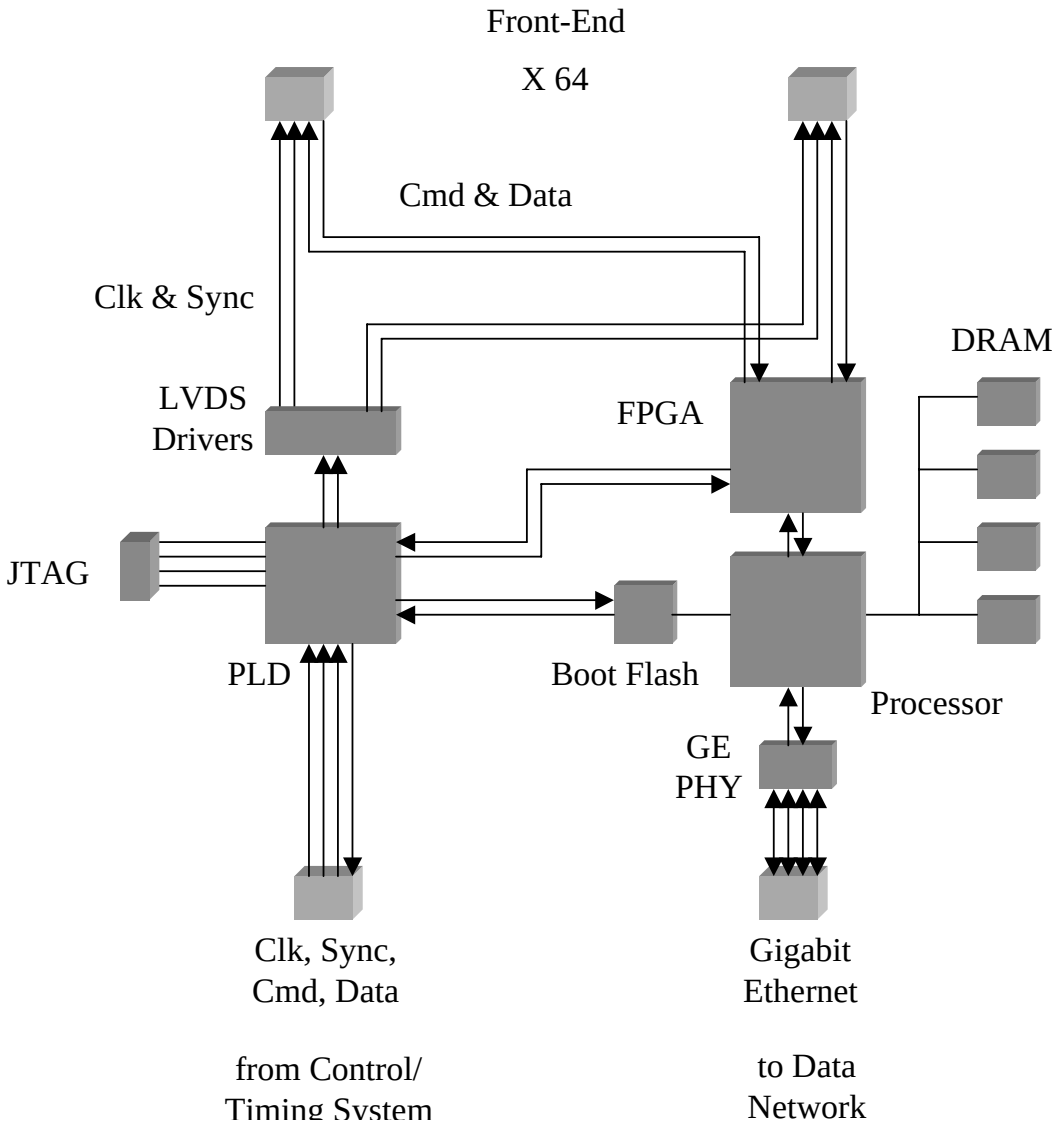

Figure 8. Data Concentrator Block Diagram

Ethernet Network

The DAQ network consists of twenty-four Gigabit Ethernet Switches (48 ports each) interconnected in a two stage shuffle (figure 9). It provides a path from every Data Concentrator to every Processor. The network can be logically subdivided so that specific Processors serve specific detector segments. This is done by simply editing the processor destination list in the Data Concentrator.

A separate router with additional security features provides connection to data storage and external networks. The total networking cost is estimated at $50K.

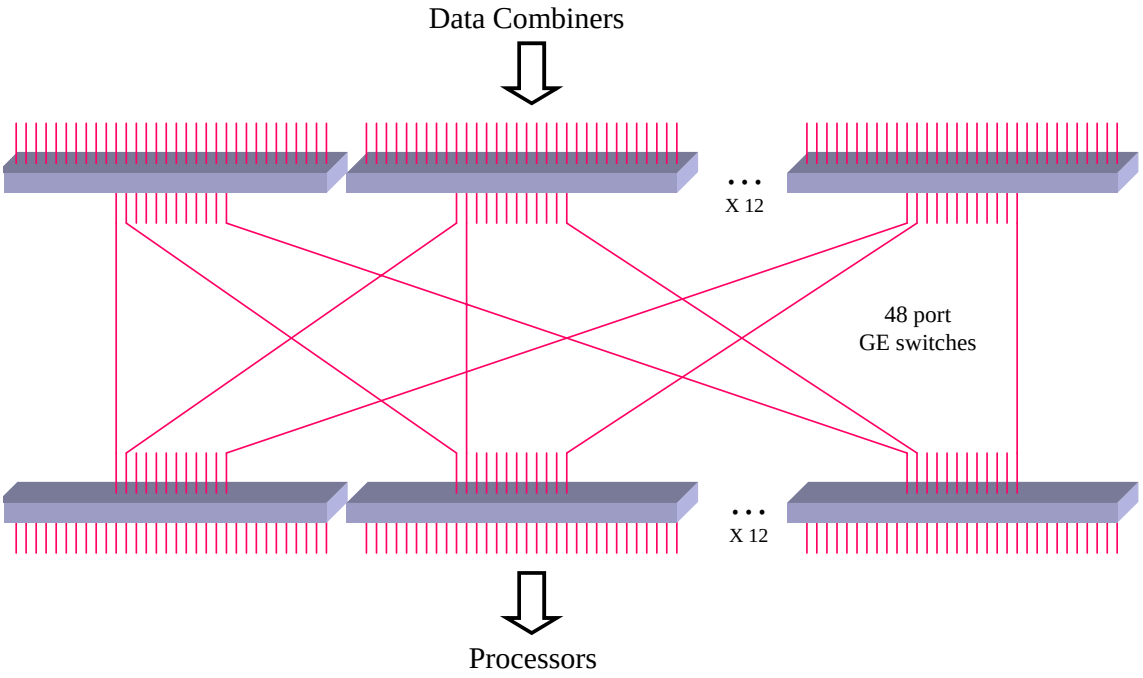

Figure 9. Gigabit Ethernet Network

Processing and Storage

The NOVA processing farm is based on commercial PCs with Gigabit Ethernet ports. The required number of processors has not yet been determined by simulation, but we will assume a total of 256 for this document.

Each processor receives data at a maximum rate of ~50 MBytes/sec. A 20 second buffer is implemented for the supernova trigger, resulting in a minimum buffer size of 1 GByte (add another 0.5 GByte for OS and application code). A basic machine meeting these requirements can be found for ~$1200, so we assume a total processor cost, with high power racks and cabling, of $350K. Dual processor machines would add another $100K at current prices, but will likely be standard at the time of actual purchase.

The required data bandwidth to storage is approximately 200KBytes/sec for spill data (30 usec/2 sec * 100 Gbps). This would fill 20 standard disk drives per year at an annual cost of $5K. With no online data compression, each supernova event requires an additional 200-300 GBytes (1 disk drive).

Copying data for a supernova event directly from processor memory to a single disk drive would take over an hour. Buffering this data in parallel to individual local processor disks and then to data storage reduces the deadtime to less than 10 seconds.

Timeslice Assembly

Data from the front-end is grouped by the Data Concentrator into timeslices of order 10 milliseconds (~1 KByte per FEB and ~64 KBytes per Data Concentrator).

Without traffic shaping, the default startup behavior would be all Data Concentrators trying to send data packets simultaneously to a single destination Processor. The internal buffering of the network can handle this for a short period of time, but the system will quickly begin to drop packets. This problem is more apparent as the timeslices become longer. With TCP/IP protocol the packets are retransmitted, and system throughput should gradually improve as Data Concentrators that are forced to retransmit lag behind and are no longer contending for the same network output ports.

Since the Data Concentrators would require buffering in any case for retransmission, the same buffers can be used to avoid blocking in the first place by using a simple traffic shaping mechanism.

All data from a single timeslice is placed in a queue assigned to a specific destination processor. Each queue can hold at least two complete timeslices at the worst-case input rate of 25 MBytes/sec. If there are 256 destination Processors, then the minimum required Data Concentrator buffer size is 128 MBytes.

Figure 10 illustrates the timeslice queuing for a simplified system with four Data Concentrators and four destination Processors. The first data timeslice (0) is placed in the first queue of each Data Concentrator, the second timeslice in the second queue (1) and so on. All queues are implemented in a common physical memory. Data is pulled from the queues in fixed size packets (~9 KBytes in the case of Gigabit Ethernet) in strict rotation. By using fixed size packets and starting the output of each Data Concentrator at a different initial queue, the system is automatically non-blocking even for very high utilization.

Figure 11 shows the output side of the network, with the data packets being reassembled in corresponding Processor queues. Timeslices from all Data Concentrators are routed to a single Processor. Timeslice queues in the Processor are deeper so that a minimum of 20 seconds of data is available when a supernova trigger is detected.

This traffic shaping algorithm in effect converts the network into a basic time-division multiplexed (TDM) switch. For the proposed system with 372 Data Concentrators and 256 Processors, there are 95,232 of these virtual channels each operating at approximately 1.5 Mbps. The latency for transferring a 10 msec timeslice from all Data Concentrators to a single Processor is approximately 500 msec.

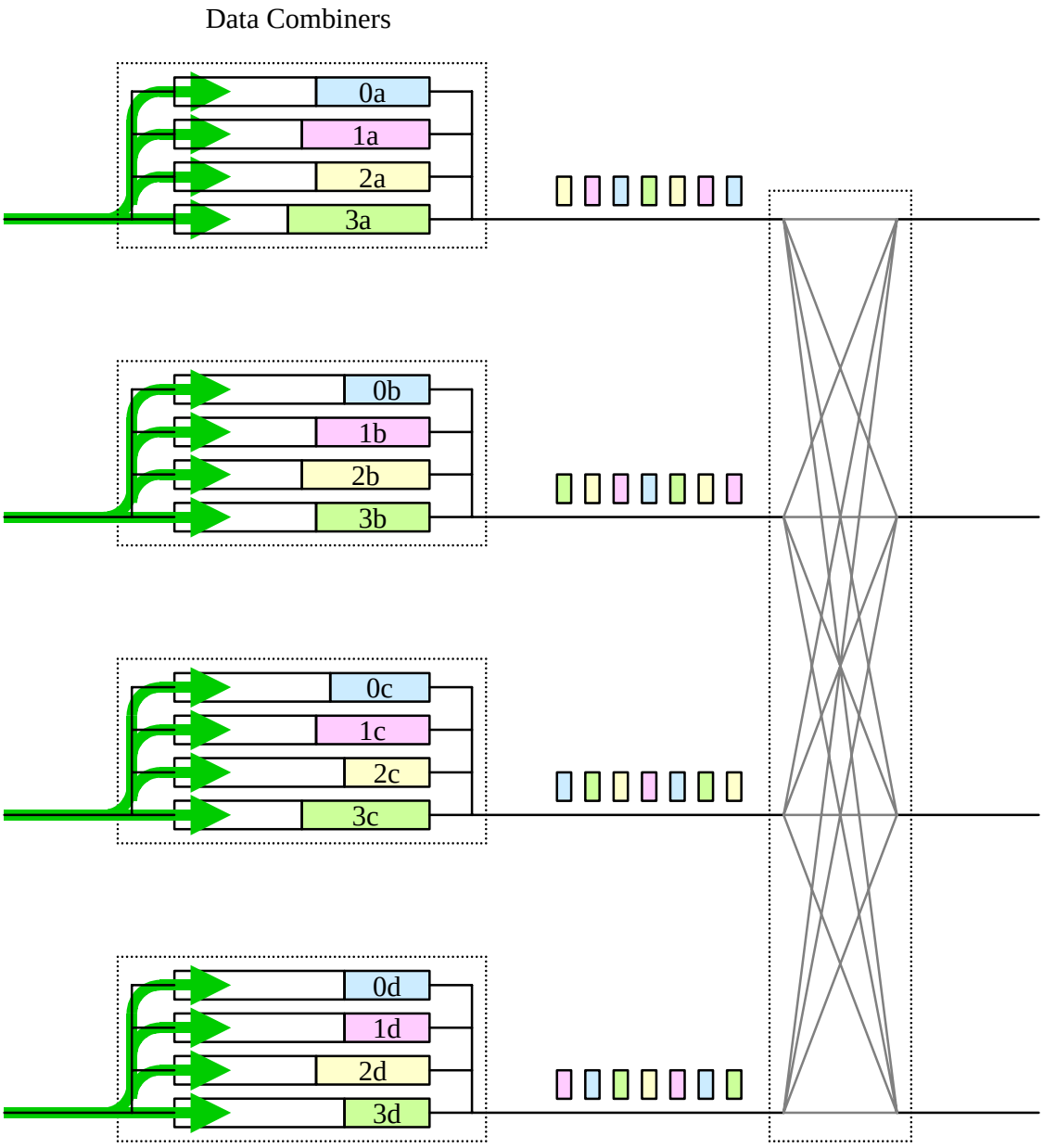

Figure 10. Timeslice Assembly (Data Concentrator)

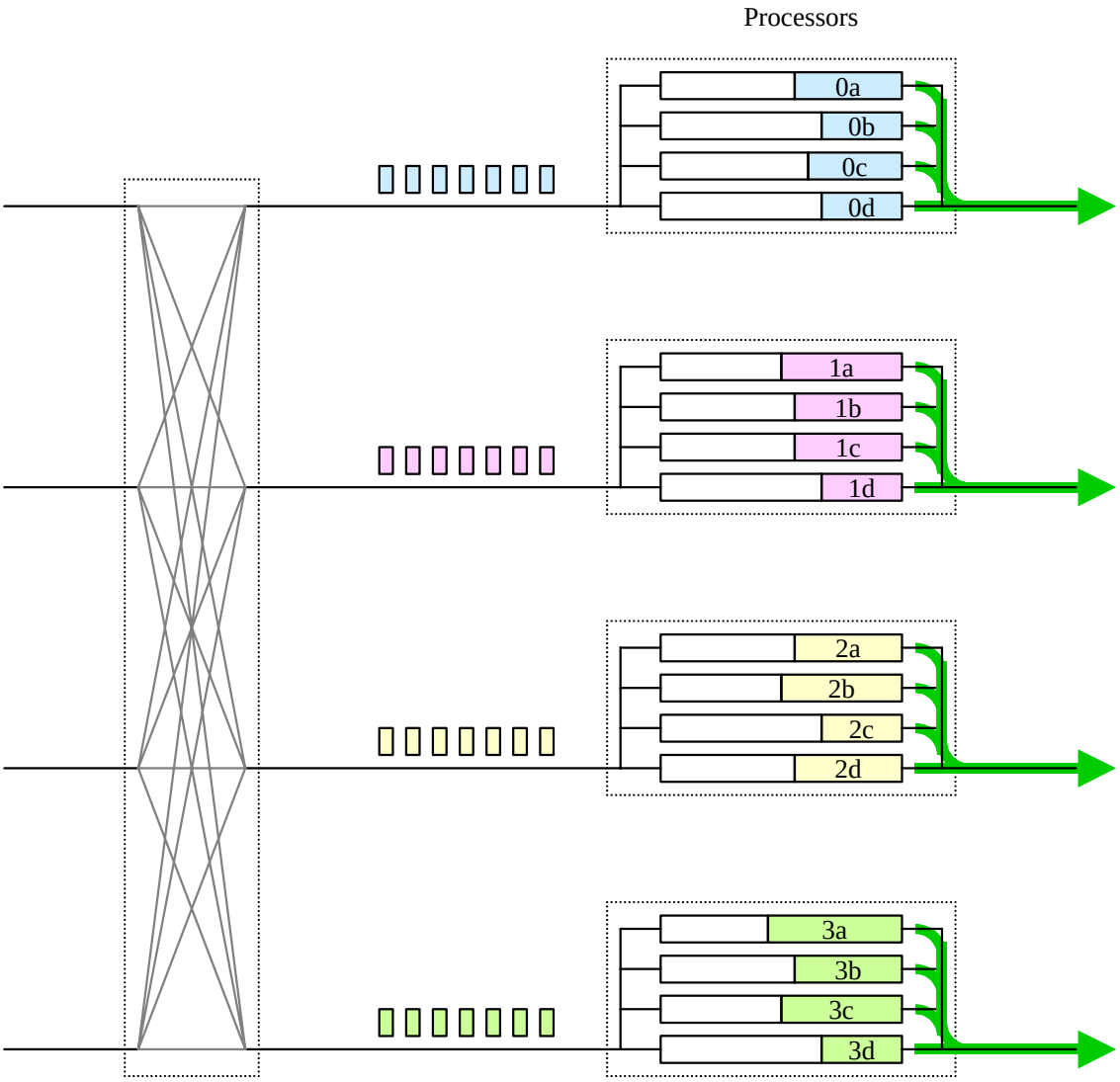

Figure 11. Timeslice Assembly (Processor)

Timing

The timing system provides a Clock and Sync signal to all Data Concentrators. The Data Concentrators retime the Sync signal for distribution to the FEBs (figure 12).

The Sync signal is a clock-synchronous reference for alignment of all Data Concentrators and FEBs. It has a defined setup and hold relative to the rising clock edge. Synchronous commands are accomplished by sending an asynchronous message over the Command link to the Data Concentrators and FEBs. The Sync signal is then used to synchronize the command to a clock edge.

The Back-end Sync (BSYNC) synchronizes the Data Concentrators. The Front-end Sync (FSYNC) is generated by the Data Concentrators at a specified clock to synchronize the FEBs.

Start Spill timing information is sent via Internet and correlated to a local GPS receiver.

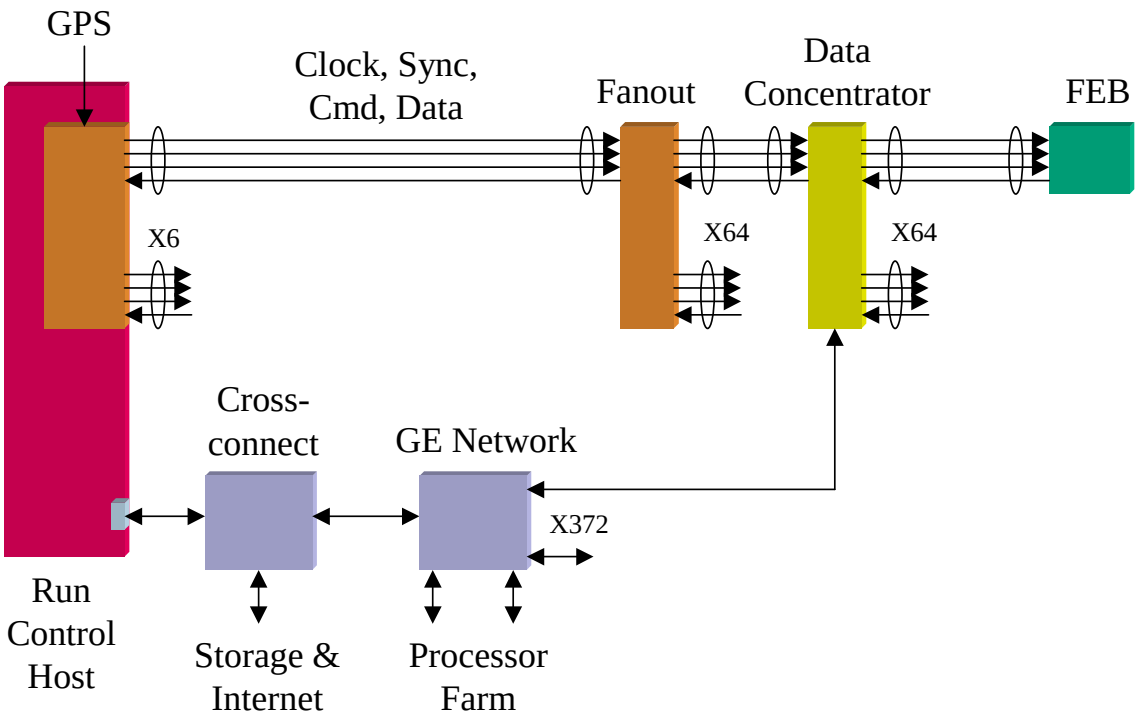

Figure 12. Timing and Control Distribution

Power Requirements

The DAQ electronics will use approximately 100KW of 120VAC electrical power. Most of this is consumed by the processor farm. The cost of AC power distribution and air conditioning for heat removal is assumed to be part of the conventional construction.

Labor

Labor associated with the DAQ hardware development is estimated at;

Management, Costing, Scheduling	0.5 FTE
Data Concentrator Design	1.0 FTE
Data Concentrator Layout and Prototype	0.5 FTE
Timing System Design	0.5 FTE
Control System Design	0.5 FTE
Timing/Control Layout and Prototype	1.0 FTE
Installation	1.0 FTE
System Test	0.5 FTE

Cost Summary

Cabling	$ 100K
Data Concentrators	$ 350K
Ethernet Network	$ 50K
Processor Farm	$ 350K
Data Storage	$ 25K
Timing System	$ 50K
Control System	$ 50K
Power Conversion	$ 25K
Labor	$ 800K
Total Base	$1800K
Contingency	$ 700K

This estimate does not include development labor for DAQ software and controls, database/controls licensing costs, or slow control hardware. It also does not yet include near detector DAQ hardware costs in excess of prototype components.